This paper has been prepared for presentation to the Federal Economic Statistics Advisory Committee (FESAC) on June 9, 2006. It represents work in progress and does not represent any agency's final positions on issues addressed. The FESAC is a Federal Advisory Committee sponsored jointly by the Bureau of Labor Statistics of the U.S. Department of Labor, and by the Bureau of Economic Analysis and the Census Bureau of the U.S. Department of Commerce.

Outliers and Influential Observations in Establishment Surveys

June 2, 2006

John L. Eltinge, Bureau of Labor Statistics
Patrick J. Cantwell, U.S. Census Bureau

Abstract

Outliers and influential observations can have an important effect on work with estimation and inference from establishment survey data. Practical development and implementation of methods to identify and account for outliers and influential observations in complex survey data require an agency to balance several factors, including: (i) the mathematical statistics properties of detection methods and prospective estimators; (ii) a range of objective functions that include traditional measures like variance and mean squared error, as well as other functions tuned to reduction of risks associated with very rare extreme observations and estimates; (iii) information available on the underlying populations of interest; (iv) cost structures; and (v) important constraints on production systems and modification thereof. This in turn has three practical implications for research on outlier methods for establishment surveys. First, the predominant outlier literature in mathematical statistics has focused primarily on area (i). Consequently, it is important to expand our mathematical structure and objective functions to account for factors (ii)-(v). Second, methods that may appear to be inefficient or inadmissible under evaluation criteria in the traditional literature may warrant serious consideration under the more complex structure defined by (i)-(v). Third, the structure defined by (i)-(v) presents an opportunity to enrich and deepen the current mathematical statistics literature on outliers.

In keeping with previous requests from FESAC, this paper focuses primarily on research that is at various stages of development, rather than on finalized research results.

Key words: Atypical observation; cost structures; downweighting; drill-down methods; M-estimation; prior information; Winsorization.

1. Introduction

1.1. Outliers in Establishment Surveys

In work with establishment sample surveys, we often encounter observations that differ substantially from most of the observations in the sample. The statistical literature refers to these observations as outliers. In a slight variation on the terminology used in Chambers (1986), we will define three distinct types of outliers: *representative outliers* which are true values that are not considered entirely unique in the population; *non-representative outliers*, which are considered to be unique in the population; and *gross measurement errors* which are outlying observations that are not true values. For example, in the study of a given industry in which most firms have less than 1% market share, one of five firms with 5-10% market share might be considered a representative outlier; a firm with 50% market share might be considered a non-respresenative outlier; and one firm's erroneous employment report of "20,000" instead of the correct "2,000" would be considered the result of a gross measurement error.

The standard survey literature on outliers tends to focus primarily on representative and non-representative outliers, and to exclude from consideration the case of gross measurement errors. The general reasoning is that gross measurement errors would have been identified and corrected in the data-editing stage, so any remaining extreme values can be considered to represent true observations. With the exception of Section 4.6 below, the current paper will follow the convention of treating all outliers as true values. For some general discussion of gross measurement errors and microdata editing procedures intended to identify such errors, see, e.g., Anderson et al. (2003), Chambers et al. (2004), Latouche and Berthelot (1992), Lawrence and McDavitt (1994), and Little and Smith (1987).

Due to the skewness of the distribution of many variables measured for businesses, outlier issues can often be important in estimation and inference from data collected through establishment sample surveys. Outliers can also be important in some household surveys, e.g., the Consumer Expenditure Survey, and other household surveys that have extreme variability in weights, extreme variability in continuous survey variables, or both. For most of this paper, however, we will exclude household surveys from further consideration.

1.2. Influential Observations

Outliers are of interest for at least two reasons. First, as implied by the terms "representative" and "non-representative" some analysts have a fundamental interest in identification of units that are so unique that they may not be of interest for certain classes of analyses.

Second, analysts often have special interest in population units that influential in the sense that their inclusion or exclusion from an estimator may lead to substantial changes in the numerical value of the resulting estimate. For the latter case, work with establishment surveys can be especially challenging because a given unit may be influential for estimation of a population aggregate quantity because: (a) it is included among the sample respondents and has a large reported value; (b) it is included among the sample respondents and has a large estimation weight; or (c) is excluded from the set of sample respondents (due either to sample selection or nonresponse) and has a very large value. For these reasons, the literature on survey outliers often

is developed in the framework of *survey-weighted influence functions.* For some formal mathematical definitions and discussion of influence functions in sample surveys, see Smith (1987), Zaslavsky et al. (2001), Gershunskaya and Huff (2004) and references cited therein.

1.3. Context for Development and Evaluation of Estimation Methods in the Presence of Influential Observations

Within the context of formal mathematical statistics, development and evaluation of outlier procedures require explicit specification of the target population and of the intended probabilistic basis for inference. These specifications in turn help one to clarify practical settings in which a proposed procedure is applicable; and to identify potential strengths and limitations of these procedures.

1.3.1. Populations and Probabilistic Foundations for Inference

In classical approaches to sample survey work (e.g., Cochran, 1977), inferential interest centers on a well-defined finite population of N units with characteristics $Y_1,\ldots,Y_N$, respectively. Probabilistic inferential statements (e.g., regarding bias, variance or confidence interval coverage rates) generally are made only with respect to the distributions induced by the sample design. We will call this Case 1.

An alternative approach views the finite population as a realization of a superpopulation model $\xi(\theta)$, and inferential interest centers on prediction of functions of the Y_i (Case 2); or on functions of the superpopulation parameter θ (Case 3). For cases 2 and 3, probabilistic statements generally are based on the superpopulation model, or on the the combined probabilistic structure induced by the design and the superpopulation model.

Another alternative views each finite population value as a sum of two terms

$$Y_i = z_i + d_i$$

where z_i and d_i are each generated by a superpopulation model, z_i represents an underlying "smooth" long-term trend in the true values Y_i and d_i represents an "irregular" or "temporary disturbance" term in Y_i. If one considers this approach, it is important to have appropriate model identification information, which generally is provided by restrictions on the model parameterization or by auxiliary data. For example, z_i may be defined through projections onto a column space of predictor variables. For this alternative, primary inferential interest may center on prediction of functions of the finite population quantities z_i, or on the parameters of a superpopulation model $\xi_z(\theta_z)$ associated with the z_i. We will call these Cases 4 and 5, respectively.

In addition, there are cases in which an analyst may wish to consider separately the properties of, respectively, the "central portion" and the "fringes" of a population. For instance, if a dominant establishment within a specified subpopulation has growth patterns or other characteristics that differ substantially from those of the other units in the population, then

inferential may focus separately on the dominant establishment and the remainder of the subpopulation, respectively; cf. the discussion of "non-respresentative outliers" in Section 1.1. In some settings, the dominant unit (and thus the two distinct subpopulations of interest) may be identified a priori, and standard approaches to subpopulation-level analysis will apply. In other settings, an analyst may prefer to define operationally the subpopulation membership through the presence or absence of outlying observations. We will refer to this as Case 6. Variations on Case 6 have been considered directly in the discussion of breakdown points in the standard literature on outliers, and indirectly in disclosure-limitation work that uses topcoding or outlier-flagged deletion of outlying values. Detailed consideration of the performance of estimation and inference procedures under Case 6 requires explicit operational definitions of the "core" and "fringe" subpopulations and associated estimands. For instance, consider a population of k-dimensional vectors $Y_1,\ldots,Y_N$ following a multivariate distribution that is reasonably approximated by a multivariate normal or other k-dimensional elliptical distribution. Then one might define the "core" subpopulation as the set of units with Y_i vectors contained in the $(1-\alpha)$ central ellipsoid of the corresponding normal distribution, and define the "fringe" subpopulation as the set of units with Y_i vectors that fall outside of this central ellipsoid.

In addition, it is of interest to detemine whether the analyst is fundamentally interested in a "core" subpopulation, rather than simply in the non-tail quantiles of the full subpopulation. In the former case, methods that adjust explicitly for outliers may be of interest. In the latter case, estimation and inference needs may be addressed through standard methods for finite-population or superpopulation quantiles, e.g., Rao et al. (1990) or Francisco and Fuller (1991).

Practical evaluation of the impact of an outlier procedure will depend heavily on which of inferential Cases 1 through 6, respectively, are of primary interest. Within the current literature, primary attention has focused on Cases 1 through 3. However, there are cases in which Cases 4 through 6 may be of serious interest. One simple example (at an aggregate level) is the use of a "core Consumer Price Index" that excludes food and energy components, which tend to be more volatile. A related example is the proposed variant on the Employment Cost Index that would exclude certain volatile forms of compensation; see Barkume and Moehrle (2004). For programs that have estimation and inference goals consistent with Cases 4 or 5, it would be of interest to evaluate explicitly the performance of a given outlier procedure with respect to the model $\xi_z(\theta_z)$ or related randomization-based properties centered on the finite-population quantities z_i.

1.3.2. Multiple Estimands

In developing methodology that will address agency needs, it also is important to note that many survey programs produce a relatively large number of estimates. This has two practical consequences for work with outliers and influential observations. First, if we consider outlier work in the context of risk management, a program manager will seek to reduce the risk of poor performance of estimators (e.g., reduce the risk that the program will publish an extreme estimate that is far from the true population value, and also is inconsistent with prior information on the likely true value) across a large set of estimands.

Second, due to resource limitations, agencies generally prefer estimation procedures that can be used for a relatively wide range of estimands, at least within a given survey program. For

work with outliers and influential observations, this preference may lead an agency to select procedures that are expected to perform relatively well across a broad set of population and design conditions.

Although these preferences are recognized in some of the survey-outlier literature (e.g., Chambers, 1996), much of the literature tends to focus on optimization for a specific estimand under a given set of conditions. Consequently, it would be of interest to expand this literature to develop additional diagnostics that: (a) allow more explicit evaluation of anticipated performance of a modified procedure, subject to specific classes of design and population conditions; and (b) produce summary evaluations of performance integrated over subspaces of the conditions considered in (a).

1.4. Summary of Material

Within the framework presented above, the remainder of this paper outlines several areas for additional research on outliers and influential observations in establishment surveys. Section 2 reviews some applicable literature and highlights some special features of complex surveys that have an important effect on the selection of outlier methods. Section 3 summarizes some specific examples of agency practice. Section 4 notes seven areas of potential research that may be especially useful for establishment survey work. Section 5 presents four related groups of questions for the Federal Economic Statistics Advisory Committee.

2. Review of Selected Literature

2.1. Sources of Influential Observations

Influential observations generally arise from one of three sources. First, as noted in Section 1, they may arise from observations that are unusually large or otherwise deviate in unusually extreme forms from the center of a reference distribution, e.g., the set of other sample values within a specified subpopulations. Second, the observation may be associated with a unit that had an unusually low selection probability, and thus an unusually high probability weight. Third, the observation may have a weight that is very large (relative to the weights of other units in the specified subpopulation) due to problems with stratum jumping; sampling of birth units or highly seasonal units; large nonresponse adjustment factors arising from unusually low response rates within a given adjustment cell; unusual calibration-weighting effects; or other factors.

The following two subsections discuss some of the issues that arise with stratum jumping and sampling of birth units. It should be noted that the situations described apply to some *but not all* Census Bureau surveys. Similarly, different surveys often apply different procedures to address the resultant outliers, or may require a different set of criteria before action is taken.

2.1.1 Stratum Jumpers

For many of the Census Bureau's economic surveys, strata are defined by the kind of business and the measure of size, that is, the size of a known or estimated characteristic closely related to those being measured in the survey. When either of these characteristics changes in a unit, its

true characteristics are no longer consistent with the sampling stratum initially assigned. In such cases, informally labeled "stratum jumpers," influential values can arise. Several types of situations can produce stratum jumpers:

(1) Acquisitions or mergers.

(2) Units that show legitimate economic growth.

(3) Units whose annual sales is not distributed as evenly across the year as other units. Examples include a new car dealer with a fleet sale in one month, or an antiques dealer with a single very large sale.

Although analysts at the Census Bureau refer many such cases to the mathematical statisticians, only a few dozen or so are treated each year in any of the surveys in the retail, wholesale, or services industries.

Under each type of situation listed above, a firm may be selected for sample in one size stratum, but its reported sales might fall into another. An initial measure of size is determined and used to place the firm into a size stratum for sampling. The measure of size is often determined from a recent economic census or survey, perhaps with some type of adjustment. As is mentioned, the firm may change in its composition following the census. For other firms absent from the census or missing the applicable data, a measure is often computed based on the relationship between the variable of interest and other available data. As an example, when a firm's sales are missing from the recent census, one can apply a regression formula to the firm's payroll, as obtained from tax records.

At times a firm's measure of size places it in a stratum where the sampling weight is quite large. After the firm is selected for sample, it may report at a much higher level, producing an influential value. This is sometimes seen with birth firms.

Stratum jumpers may be treated in different ways in different surveys. For example, suppose a firm is classified in one kind of business or industry erroneously. Later, the mistake is detected. If the firm is a certainty (that is, selected into sample with probability 1), it is generally re-classified. However, in certain industries if it is a noncertainty single-unit establishment, it retains the classification assigned at the time of sampling. It continues to report or have its value imputed so that it is representative of the units within that industry.

Occasionally a firm is classified in one kind of business, but is doing business in two or more areas within the same industry. An example is a firm initially classified as a convenience store, but which also sells gasoline. Later, its sales (revenues, inventories) grow in the secondary kind of business. If the firm is a single-unit establishment, it is not re-classified. However, for multi-unit firms, separate reporting parts may be established for the several kinds of business. This action can alleviate potential problems of (i) inconsistency between reported values and the sampling weight, or (ii) misrepresentation of the kind of business activity.

2.1.2. Sampling of Birth Units and Highly Seasonal Businesses

An influential value can arise when a frame unit is assigned a measure of size for sampling that is considerably smaller than the corresponding value the unit later reports when canvassed in the survey. The initial measure of size can cause the unit to be placed in a sampling stratum with smaller units. Under optimal allocation, a smaller proportion of units is typically selected from such a stratum of smaller units, so that the resulting sampling interval and sampling weights are larger than what would have been applied if the measure of size reflected the larger value that applied to the establishment later.

This situation may occur relatively more frequently in birth sampling. After the initial selection of units for a survey, new firms may be created, and some firms may be restructured through business or legal transactions such as mergers. In some surveys, such as the Census Bureau's monthly and annual surveys of retail and wholesale trade, births are added during the five-year life of the sample.

To simplify the description a bit, birth processing uses two stages of sampling. To start, the Census Bureau receives administrative records indicating that a new or restructured firm is active in the industry. At this point, the available data provide a rough indication of industry classification and a recent measure of the firm's payroll or expected number of employees. Based on this information, the unit is assigned to a stratum according to its industry classification and an initial measure of size, and a first-stage sample is chosen. Selected units are asked for more precise information on their industry, and two recent months of sales or receipts. This information can then be used to better classify the unit and place it in the proper size stratum. A second stage of sampling is then applied to the units for inclusion in the regular monthly or annual survey.

Although birth sampling helps maintain coverage of the evolving frame, it can introduce problems that may lead to influential values in the data. First, the measure of size for a birth generally is not as accurate as for a firm selected when a new sample is introduced. For the latter, the measure is often determined from the same variable captured at the time of the most recent economic census, and is considered to be quite accurate. A birth, however, was not active in business during the census, but is a new or restructured business. As the Census Bureau tries to obtain size and classification information as early as possible in the life of the new firm, the reporting in some cases does not reflect the size to which it will eventually grow. The problem is one of trying--at an early stage, and with limited information--to predict the firm’s ultimate size so as to place it in the most appropriate size stratum. In addition, some units in the first stage of the birth sample are active but do not respond to the questionnaire. Their measures of size for the second stage must be imputed using administrative data, and may be inaccurate.

A second problem can arise because we apply a two-stage design when sampling births. The unit's weight is the product of the inverses of the probabilities of selection at each stage. We may learn from the first stage of sampling that our initial measure of size was too large, and that the unit should have been placed in a smaller size stratum, where it would have been subjected to a smaller probability of selection. That is not a serious problem, as we can adjust downward the probability of selection at the second stage so that the product reflects the appropriate probability corresponding to the more accurate measure of size.

However, occasionally the first stage of sampling tells us that the birth unit’s sales or revenues appear to be somewhat *greater* than we had predicted using the initially available payroll or employment data. Based on that information, we had placed the unit in a stratum with mostly smaller firms. For such a unit, the weight may already be too large before the second

stage. If we want to retain an unbiased selection procedure, we cannot use the second stage of sampling to adjust the probability and weight to the correct level, as it can only make the probability smaller--and the weight larger. When this happens, the unit has a weight that is larger than that corresponding to its size. If this firm's reported sales or receipts later increase dramatically, the weighted value may produce an influential observation that must be addressed.

Similar issues arise in the sampling of some highly seasonal businesses. For example, in the Current Employment Statistics program of the BLS, the sample design is stratified by state, industry and size, where size is defined by employment count recorded in the Quarterly Census of Employment and Wages in March of a baseline year. For some businesses with highly seasonal employment patterns, an establishment may have a low employment count in March, and thus have a low probability of selection and a correspondingly high probability weight. The same establishment, however, may have a very large seasonal relative increase in employment between March and July. The combination of a large weight and a large relative increase in employment produces a highly influential observation.

2.2. Downweighting

In many establishment surveys, when an outlying or otherwise influential observation is identified it is often standard practice to reduce the survey weight associated with that observation. For some general discussion of downweighting procedures, and the practical implications of such downweighting, see, e.g., Chambers (1996), Detlefsen (1992), Elliott and Little (2000), Potter (1988, 1993), Theberge (2000) and Zaslavsky et al. (2001).

A relatively common practice is to set the survey weight equal to one. In an informal sense, this would be consistent with the idea that the identified outlier is a "non-representative" outlier. Such units sometimes are called "atypical." In addition, in some cases setting a weight equal to one may be viewed as the limiting case of more refined adjustment procedure like Winsorization or M-estimation, as discussed below.

2.3. Winsorization and Other Modifications of Reported Microdata

The general idea of Winsorization is that if an observation exceeds a prespecified cutoff value c, then the observation will be replaced by that value c. One common example of this is "topcoding" of certain values, often in conjunction with disclosure-limitation procedures. In a more formal framework following Chambers et al. (2000), one may define two types of Winsorization. Under Type I Winsorization, we consider original observations Y_i and an estimator of a population total $\hat{T}_W = \sum_{i \in s} w_i Y_i^*$ where $Y_i^* = c$ if $Y_i > c$ and $Y_i^* = Y_i$ otherwise. Under Type II Winsorization, the same estimator of a population total is used, but we define $Y_i^* = (Y_i / w_i) + c(w_i - 1)/w_i$ if $Y_i > c$ and $Y_i^* = Y_i$ otherwise.

For some general background on Winsorization, see, e.g., Ernst (1980), Fuller (1991), Kokic and Bell (1994), Rivest and Hurtubise (1995), and Searls (1966). In addition, Wolter (1996 a, b) discussed the possible use of Winsorization for the Current Employment Statistics program of the BLS.

2.4. M-Estimation and Robust Generalized Regression Estimation

In the past thirty years, there has been extensive development of generalized regression estimation methods in which weights are modified to incorporate information from auxiliary variables. More recently, generalized regression estimation methods have been linked with the literature on M-estimation that developed originally in non-survey contexts. For a general review of these methods, see, e.g., Beaumont (2004), Beaumont and Alavi (2004), Hulliger (1995), Theberge (2000), Welsh and Ronchetti (1998) and references cited therein.

2.5. Multivariate Issues

Both establishment and household survey data tend to be multivariate in the sense that we generally collect several different key measurements from our sample respondents. In most cases, two or more of these measurements have skewed or heavy-tailed distributions, and thus may be subject to outlier issues. For these cases, agency practice commonly has been to apply outlier methods to one or more of these key measurements, and possibly to specified functions of the measurement vector (e.g., the ratio of two measurements). This approach has the advantage of simplicity, but may lead to several complications. First, it is possible for a vector to be extreme (as measured by Mahalanobis distance or similar multivariate measures of distance from the current distribution) and yet have each of its univariate components within customary tolerance bounds. See, e.g., Beguin and Hulliger (2004), Franklin et al. (2000), and references cited therein.

Second, if the univariate components must satisfy certain exact or approximate functional constraints (e.g., additivity constraints), then direct use of Winsorization or other customary modifications to the observations may be problematic. For this reason, some authors (e.g., Chambers, 1996) note that weight reduction can provide approaches to multivariate outliers that are relatively simple operationally. The latter approach, however, leads to the downweighting of univariate values that are well within tolerance bounds. An alternative approach used by some agency programs is to apply outlier tools to only one critical measurement or derived value. For example, the BLS Survey Occupational Illnesses and Injuries (SOII) applies a simple tolerance interval approach to an incidence-rate ratio defined to equal the number of injuries divided by the number of work-hours reported for a given sample establishment. For establishments with extreme ratios, weights are set equal to one. Note that if the denominator term is small, then the abovementioned ratio may be large even if the numerator term is well within univariate tolerance limits.

3. Examples of Agency Procedures for Outliers and Influential Observations

The current paper is focused primarily on identification of classes of methodological research questions of interest to statistical agencies that work with outliers and influential observations in establishment surveys. In the interest of space, we have not attempted to provide a detailed catalogue of specific outlier and influential-observation methods currently in use at the

Census Bureau and the Bureau of Labor Statistics. For some background on specific methods used at the Census Bureau, see Bienias (1995), Bienias et al. (1994), Detlefsen (1992), and Gomish and Sigman (2003). In addition, Scott et al. (1999) provide a detailed discussion of outlier and influential-observation methods at the BLS, and U.S. Department of Labor (2003), Barsky, Dorfman, Dworak-Fisher and Ernst (2003), Barsky, Dorfman, Dworak-Fisher, Ernst and Guciardo (2003), Gershunskaya and Huff (2004) discuss specific cases.

The following examples are intended to illustrate some of the practical issues encountered in specific Census Bureau surveys, but do not cover the full range of outlier procedures used at the Census Bureau. Similar comments apply to some establishment surveys at the Bureau of Labor Statistics, but are excluded in the interest of space.

3.1. Influential Observations Arising from Three Levels of Sampling in the Commodity Flow Survey

The Commodity Flow Survey (CFS) is conducted by the U.S. Census Bureau in partnership with the Bureau of Transportation Statistics. The CFS selects a sample of shipments; the key characteristics are the origin, destination, value, and weight of the shipment. Two aspects of the survey design allow for influential values.

First, under the three-stage sampling design, the sampling weights are quite large for some shipments; in fact, the survey design does not constrain the maximum sampling weight that any shipment may have. The Census Bureau selects a sample of establishments, that is, business locations. In each quarter of the reference year, the establishment is assigned a specific week. The respondent is asked to select a sample of no more than 40 shipments from all those made during the given week. A maximum weight is imposed on the first stage of sampling, the selection of establishments, while the second-stage weight is 13. But in the third stage, the selection of shipments, the population of shipments for the week can theoretically be unlimited. Large third-stage weights have been observed. Multiplying the weights from the three stages can produce some extremely large weights. In the 2002 CFS, the largest final weight for a shipment was close to 500,000.

A second reason for potential outlying values in the CFS is the variability of the characteristics measured. The value of a single shipment ranges from almost nothing to millions of dollars. Although a firm typically does not ship thousands of items worth millions of dollars each, the survey responses are subject to the randomness of what falls in sample and, sometimes, to the initiative of the responding sampler. It appears that some respondents have tried to "help" us take the survey by including the shipments that they think are representative of their usual business patterns, or that we might otherwise miss in the sample. To address these problems with (1) the true variability of shipment value or size and (2) our dependence on the respondent in selecting their sample of shipments, the Census Bureau created a certainty stratum of shipments when redesigning the survey for 1997. A question was added: "In the last three months did this location have any individual shipments with a value over $2,000,000?" If the response "Yes" was checked, that establishment was contacted by telephone to solicit all such large shipments. Still, sampling and nonsampling error can creep into the process.

The concerns brought about by outlying values in the CFS are different from most of those we have discussed. Unlike the monthly and annual surveys, in which measures of change are typically paramount, CFS is conducted every five years, so that measures of level may be

more important. But detecting and addressing outliers can be similar. Even with all the editing applied to the shipments in sample, it is not uncommon to notice an anomaly at the macro-level. For example, one might observe that most of the U.S. total of some commodity was shipped from Montana to Florida, simply because of an extremely high weight on the shipments from one establishment in Montana. Similarly, the average length of shipments sent via a specialized transportation mode might be around 3,000 miles, because the only *sampled* establishment shipping via that mode was on one coast, sending all of its goods to the other coast. In such cases, an unexpected macro-level total leads the analysts to examine the micro-level data.

To summarize, influential values in the CFS arise due to a combination of factors:

(1) When designing the sample, the available data are limited to the *industry* and location of each establishment in the frame, and its annual value of total shipments.

(2) The tabulation cells are based on the *commodity transported* (rather than the industry of the shipper), the mode of transport, and the origin and destination of the shipments; and measure such characteristics as total dollar value, weight, distance shipped, etc.

(3) While the populations underlying the tabulations in (2) correspond to dozens of dimensions, the sampling based on (1) uses only three dimensions.

(4) The populations in (2) are themselves skewed in a variety of ways not reflected in the sampling information in (1); the analysts cannot possibly know the relationships for the hundreds of thousands of cells published.

(5) As described above, the three stages of sampling can introduce very large weights.

For a survey like the CFS, from which tabulations are made in up to four dimensions, even the concept of an influential value is hard to define. What might be considered influential or an outlier at one level of tabulation might not be at another.

3.2 A Winsorization Procedure for Outliers in the Census Bureau's Survey of Construction

In the Census Bureau's Survey of Construction, a Winsorizing outlier procedure has been in place for a number of years to help stabilize the characteristics sales price and contract value. When the weighted value from a respondent is greater than a specified cutoff, that value is reduced so that the adjusted weighted value is equal to the cutoff.

The cutoff was determined so that the mean squared error (MSE) of the estimate is minimized. Here, the MSE is computed as the sum of an estimated "simplified" variance and the square of the bias. The simplified variance is the product of a design effect multiplied by the variance computed as if the sample units were selected as one large sample with unequal probabilities. The bias is estimated as the difference between the characteristic estimate with and without the outlier adjustment. Various cutoffs were tested until the one that minimizes the MSE was found.

Over time a problem had developed with this Winsorizing outlier procedure: the values of the responses had increased. With a constant cutoff for outliers, more and more outliers were identified and their values reduced, producing a significant estimated bias in the characteristic estimates and a MSE that was no longer minimal with respect to the cutoff. To address the problem, methodologists increased the cutoff value, thereby accommodating the inflation of response values. Further, cutoff values for past data were re-computed and used to adjust historical series of estimates.

3.3 Addressing Outliers in the Consumer Expenditure Survey and the Survey of Residential Alterations and Repairs

In the Consumer Expenditure Survey (CE) and the Survey of Residential Alterations and Repairs (SORAR), the key variable is expenditures, which can be classified as additions, alterations, additions and alterations to outside structures, and maintenance and repairs. (Prior to 2004, a fifth category was separated, major replacements; it is now included in the remaining categories.) While the data for both surveys are collected by the Census Bureau, the Bureau of Labor Statistics sponsors the CE and publishes its tabulations.

Gomish and Sigman (2003) describe outlier procedures for these two surveys. Detected outliers are Winsorized, that is, the value of the response is truncated. Specifically, if the weighted value from a household is greater than a fixed cutoff percentage of the total, then the weighted value is reduced to the cutoff level.

At one time, these surveys used a different Winsorization method to treat outliers, in which the weighted threshold for expenditures was a fixed cutoff, $300 million--rather than a percent of the weighted total. Weighted values that exceeded $300 million were reduced to this cutoff. Over time, as values increased, this fixed cutoff was detecting a greater number of outliers, to the point where the downward bias in the estimates of total introduced by the procedure was considered to be too large. Applying a fixed percentage cutoff would allow the Winsorization technique to automatically conform to the rising values of expenditures, without the need to change the cutoff value itself. Research was conducted to determine the optimal percentage cutoff, one that would minimize the mean squared error as estimated from the data.

The Winsorization approach has been adjusted further. An iterative procedure re-sets the total--and thus the threshold amount--as each new outlier is confirmed and reduced in value. In addition, an upper bound is placed on the absolute relative bias introduced by the procedure. If the percentage cutoff constant would induce so many outliers that this bias bound were exceeded, the percentage cutoff is increased and fewer outliers are identified.

An important procedural issue is the level at which to apply the outlier technique, that is, total expenditures, individual types of expenditures, or somewhere in between. If the technique is applied at the individual level, different percentage cutoffs would be preferred for different types. Research implied that addressing outliers at this low level would overly complicate the process and would detect more outliers than would be desired. Thus, the procedure is applied at the aggregate level with a 3% of total cutoff for the CE survey, and a 4% of total cutoff for SORAR. Outlying values are detected and reduced before being inserted into subtotals for the various types and combinations of types.

It should be noted that *unweighted* extreme values of expenditures are seldom changed by the editing process. Therefore, such values--whether extremely large or small--are later isolated using resistant fences. Respondents are then called to verify that the data are accurate.

3.4 Addressing Outliers in the Census Bureau's Annual Capital Expenditures Survey

In the Annual Capital Expenditures Survey, methodologists hoped to take advantage of several variables when trying to identify outliers. Detlefsen (1992) provides a general description of the resulting procedure, based on methods described in in Bibby and Toutenberg (1977, Chapter 2).

Data are standardized, the dominant principal components are determined, and the Mahalanobis distance computed and analyzed to see which points fall far from the standardized center. Suspected outliers are then sent to analysts to check the accuracy of the responses. Unresolved cases are returned to the methodologists, who check other criteria, including

(1) is the weighted response at least 20% of the estimate of total?

(2) is the weighted response at least five times that of the next largest response in the same sampling stratum?

(3) is the weight large enough to justify a change?

After analysis, if the case is considered to be an outlier, its weight is reduced by applying what one might refer to as the Bibby and Toutenberg approach (BTA). Under this approach, one reduces the weight of an outlying value by an amount that seeks to produce the minimum mean squared error (MSE) of the estimator.

To describe the BTA in slightly more detail, suppose that, based on recent historical data, the sample design and estimation procedure produce an unbiased estimator (say, t_{unb}) with a coefficient of variation, *CV*. A reduced estimator, $r\, t_{unb}$, can be computed that minimizes the MSE, where *r* is a constant. It is easy to show that the minimizing *r* is $1/(1 + CV^2)$. One can simply multiply the unbiased estimator by *r*, or, equivalently, can change the weight of an observation, y_i, from w_i to $(r\text{-}1)\, t_{unb}/x_i + w_i$.

In the Annual Capital Expenditures Survey, when an outlier is identified and the value confirmed as accurate, its weight is reduced by applying the BTA. An advantage of this procedure is in retaining the true response value, rather than suppressing it and imputing. Further, the reduced weight alleviates the effect and the possibility that the case might not be truly representative of w_i-1 other cases.

However, there are disadvantages as well. The BTA was not developed as a technique to address outliers. It was meant to produce a better estimate of total under the sample design by balancing the bias-variance trade-off, that is, to work on t_{unb}, rather than on individual observations. Thus, the procedure can produce surprising or undesirable results for the estimator or the outlier. For one, when the coefficient of variation of the estimator is very small--as is often the case for important statistics--the weight of the outlier tends to be reduced only slightly, perhaps no more than 10 or 15 percent.

To look at this another way, suppose that the coefficient of variation based on historical data remains the same (*CV*), and that the responses of all values *except* for one outlier remain the same. Then if one lets the outlying value increase, the revised weight of the outlier *increases* (relative to the revised weight from a smaller outlying value). As the outlying value increases without bound, the revised weight approaches $r\ w_i$. So, even for outlying values that are not extremely high, the revised weight may easily be greater than 90% of the original weight, depending on the value of *CV*. That is, because the original intent of the TBA is to reduce the estimator by what is often only a small amount, the weight and thus the weighted value of the outlier might not be reduced as much as would be desired.

3.5 Addressing Outliers in the Census Bureau's Monthly Retail Trade Survey

Before 1999, the Census Bureau primarily used two edits in the Monthly Retail Trade Survey to detect outliers. These procedures identified two sets of cases: one set for analyst review, and a second set to be automatically suppressed and imputed. Problems with each edit developed, which caused the Bureau to examine and implement the Hidiroglou-Berthelot edit, and eliminate the edits used earlier. For a detailed discussion of application of this approach to the Monthly Retail Trade Survey, see Hunt et al. (1999).

The first edit measured the firm's weighted share of the market for the current month, and compared it to that for the prior month. If the computed statistic was in a specific range of large values, the case was sent to analysts for review; if the statistic was larger than the endpoint of that range, the value of sales was suppressed and imputed.

Problems with this edit had developed. The edit had been put in place when the sampling units were single-unit establishments. Large changes were required to indicate a potential outlier. However, in the 1990's, the Census Bureau stopped providing geographic detail in the monthly trade surveys, and began selecting sample units at the level of the company or tax (employer identification) number, which often comprises many establishments. For companies with many establishments, large but reasonable changes in market share from month to month were being identified as outlying responses. Often, after review by analysts, such responses were verified to be accurate, and the values retained in the estimate. However, because of the processing flow, they were taken out of the imputation database. That is, their values did not contribute to the imputation for nonresponding or truly inaccurate reports.

The second edit computed the ratio of sales for the current and prior months, and compared it to the ratio for the entire kind of business. If the ratio for a specific firm was too small or too large, the value of sales for the current month was suppressed and imputed. This procedure treated firms with large and small dollar volume similarly, even though small firms have little effect on the estimate of total within the kind of business.

Methodologists from the Census Bureau studied the use of the Hidiroglou-Berthelot (HB) edit on past data from the Monthly Retail Trade Survey. This procedure found cases that were later confirmed or suspected to be true outliers, and others that the current edits had missed. Further, the HB edit did not identify as many firms whose data--upon further analysis--appeared to be good. In January, 1999, the HB edit was implemented in production in conjunction with the (then) currently used edits. After a test period of three months, the current edits were eliminated in April, 1999.

In addition, a study currently led by M. Mulry is exploring ways in which to improve the treatment of influential observations in the estimation of total revenue from the Monthly Retail Trade Survey. An observation is considered influential if the estimate of total monthly revenue is dominated by its weighted contribution (Chambers et al 2000). Influential observations occur infrequently but are problematic when they do appear. In keeping with the restrictions given in Section 1.1, this study assumes that the influential observation is true although unusual, and not the result of a reporting or recording error. The goal is to find methodology that uses the observation but in a manner that assures its contribution does not dominate the total.

The current MRTS processing uses the Hidiroglou-Berthelot algorithm (1986) each month to create the imputation pool (Hunt, Johnson and King, 1999). The Hidiroglou-Berthelot algorithm designates observations that should be reviewed and sometimes suppressed from the imputation database. The intent is for the treatment of influential values to complement, not replace, the Hidiroglou-Berthelot algorithm.

Per the discussion in Sections 2.2 and 2.3 above, two general approaches are available for the treatment of influential observations in estimation: (1) trimming the weight, sometimes called constraining the weight, and (2) modifying the outlying observation so that it has less impact on the estimate of the total. If the business will continue to report influential values, then possibly it could be considered as belonging to another sampling stratum, and a change in the weight may be the better option. If the influential value appears to be a rare occurrence for the business, then adjusting the value may be more desirable.

The basic strategy has been to identify candidate methodologies, perform a preliminary study with actual data for a month that contained an influential value, and then conduct a simulation to further evaluate the methodologies that demonstrated promise in the preliminary study. The evaluation criteria include the number of influential observations that are detected, including the number of true and false detections made. The evaluation criteria also include estimates of bias, variance, and mean squared error. In addition, the evaluation will include an assessment of the impact on measures of change, particularly the month-to-month ratio of revenue.

The preliminary study examined weight trimming and methods that modified the influential observation. The weight trimming approach made the arbitrary choice of cutting the weight to one-third of what it was originally. The methods examined for modifying the outlying observation were as follows:

- Winsorization (Chambers et al. 2000)
  - specifying a cut-off value for observations by stratum (Kokic and Bell 1994)
  - specifying an individual cut-off value for each observation (Clarke 1995)
- Reverse calibration (Chambers and Ren 2004)
- Generalized M-estimation (Beaumont and Alavi 2004, Beaumont 2004)

From the initial results, reverse calibration and Winsorized cut-off by stratum did not show promise. The data appeared too extreme for reverse calibration, and the Winsorized cut-off by stratum identified too many false outliers.

The simulation study is examining Winsorization with an individual cut-off for each observation and the generalized M-estimation. Jean-Francois Beaumont of Statistics Canada has sent software he developed for his research. The software provides the choice of modifying the influential observation or modifying its weight. Although modifying the influential observation is the most attractive choice for a one-time occurrence, the option of modifying the weight may

prove helpful in deciding how to weight a business that appears as though it will continue to be influential.

4. Research Issues in the Development and Implementation of Methods for Outliers and Influential Observations

Sections 1 and 2 noted briefly several areas in which further research would be warranted. This section outlines seven additional research areas that would be of practical interest to government statistical agencies.

4.1. Variance Estimation, MSE Evaluation and Inference Methods that Account Explicitly for Outlier Adjustments

Section 2 noted that outlier adjustment methods are intended to produce an estimator with reduced variance, but at the price of inducing some bias in that point estimator. This leads to three issues related to variance estimation and inference. First, if a standard variance-estimation procedure is applied to survey microdata after application of downweighting, Winsorization or other adjustments, then the use of customary variance estimators may not fully reflect all components of variability. In addition, due to the abovementioned bias, a customary table of standard errors generally will not fully reflect the overall estimation error. Consequently, for cases in which outlier-adjustment methods lead to a nontrivial bias (relative to the standard error), it may be preferable to report estimates of the square root of the estimated mean squared error, instead of the customary standard errors. Third, for cases in which formal inference (e.g., the construction of an approximate 95% confidence interval) is of interest, one would need to adjust the inference procedure to account for bias in the adjusted point estimator.

4.2. Evaluation and Reduction of Risks Not Fully Reflected in Mean Squared Error

In keeping with Section 1.3.2, the concerns of some program managers regarding outliers tend to center on the risk that one or more influential observations may lead to publication of estimates that are extreme relative to estimates published for previous time periods, or relative to an informal Bayesian prior distribution for the underlying true population values. For a given estimand, the risks of primary concern may involve relatively low probabilities (e.g., differences between the published estimate and the true value greater than or equal three standard errors), and thus have a relatively minor impact on the mean squared error of a given estimator. However, because the program manager must consider these risks for a relatively large number of estimands, the cumulative risk (summed over the estimands of interest) may be nontrivial. Consequently, it would be of useful to develop tools to characterize and quantify these risks for individual estimands and groups of estimands; and to develop estimators intended to minimize the abovementioned cumulative risks within specified groups of estimands.

4.3. Conditions Under Which Adjustment for Outliers and Influential Observations May Be Justified

In work with outliers, agency needs and risk profiles sometimes are not fully reflected in the bias-variance trade-offs and mean squared error evaluations that have dominated the formal mathematical statistics literature on outliers in surveys. For instance, consider again the framework defined by Cases 1 through 6 in Section 1.3.1. If inferential interest centers on one or more of Cases 1 through 3, then removal or downweighting of influential observations involves a risk of "oversmoothing." Stated in informal terms, application of an outlier procedure may "smooth out" some of the most interesting new developments in a given sector of the economy. On the other hand, under Cases 4 through 6, outlier adjustments may be appropriate, provided the data users have a clear understanding of the "smoothed" target population under consideration, as well as the possible biases involved in the possible use of smoothed estimators for inference under Cases 1 through 3; cf. the discussion of multiple estimands in Section 1.3.2.

Also, in keeping with statistical agency traditions of objectivity, transparency and reproducibility, there is a general preference for outlier procedures that are prespecified and require relatively little judgment by individual data analysts. For most large-scale surveys, it is unrealistic to expect entirely "automated" outlier procedures, due to the complexity of the economic, health or social variables being measured. However, public confidence in the objectivity, transparency and reproducibility of reported results can be enhanced by careful attention to prespecification of outlier procedures, clarification of the target populations, and quantification of the associated error components and inferential risks.

4.4. "Drill-Down" Methods: Accounting for Analyst Information and the Cost of Data Review

With some exceptions, the formal statistical literature on outliers and influential observations tends to use (implicitly or explicitly) two important assumptions.

(i) Examination of sample observations to determine outlier status can be carried out at costs that are trivial relative to other cost components, e.g., for data collection.

(ii) The reference distributions used to identify outliers or influential observations also can be determined at a relatively low cost.

For some establishment surveys, however, assumptions (i) and (ii) may not hold. In addition, for some cases the reference distributions of interest in (ii) may be quite complex and depend heavily on the subpopulation under consideration; cf. Section 1.3.2. For such cases, it may be efficient to consider the use of "drill down" procedures that involve two distinct steps.

(a) For each of many prespecified "estimation cells" (often defined by the intersection of simple classificatory variables like geography, establishment size, industry and occupation) the analyst reviews preliminary cell-level point estimates and identifies cells that appear to be "extreme."

(b) Within cells identified as "extreme," the analyst examines individual microdata records to determine whether the "extreme" cell-level estimate can be attributed to a small number of outlying or influential observations.

It appears that procedures following (a)-(b) are used in several statistical agencies. See, e.g., Luzi and Pallara (1999) and DiZio et al. (2005). Because "drill down" procedures may rely more heavily on analyst judgement, some surveys have moved away from "drill down" procedures. For example, the Survey of Occupational Illnesses and Injuries, a federal-state cooperative program of the Bureau of Labor Statistics, has in recent years moved states away from drill-down procedures and toward procedures that are more consistent with the predominant statistical literature.

However, the formal literature in mathematical statistics has not considered (a)-(b) in depth, and it would be worthwhile to do so. Some appropriate questions would be:

- To what extent can one characterize step (a) as an analyst's comparison of the cell-level estimates to an (implicit) Bayesian prior distribution determined through auxiliary information like overall patterns observed at higher levels of aggregation, survey results for the same cells in previous periods (for repeated surveys), and "local knowledge" of specific economic events like strikes or plant closings.

- Similarly, to what extent can one characterize step (b) as an analyst's comparison of establishment-level observations to an implicit Bayesian prior distribution based on the other observations recorded in that cell, and possibly on other sources of information like previous-period results?

- Under what cost structures (i.e., the relative costs of collecting an observation, reviewing a cell-level estimate, and reviewing microdata within a given cell) is a "drill down" procedure more efficient that the procedures developed under conditions (i)-(ii)? For this question, the evaluation of efficiency would need to account for the abovementioned cost structures and standard measures of estimator performance like mean squared error. In particular, the evaluation of mean squared error would involve multiple steps and would reflect the risks of Type I and Type II error in the initial screening step (i). For example, if outliers are distributed randomly across cells, then aggregation effects would tend to mask some outliers in step (a). On the other hand, if outlying true values arise according to a hierarchical model that has a structure that matches the cell-based structure in (i), then outliers will tend to appear much more frequently in specified cells, and so procedure (a)-(b) may be relatively efficient for identification of outliers.

- To what extent can an agency use Bayesian or frequentist methods to "automate" the drill-down procedure (a)-(b), and thereby characterize the trade-offs among cost components and mean squared error for a "drill-down" procedure, relative to a procedure developed under conditions (i)-(ii)?

4.5. Tools for Comparison of a Proposed Methodology with a Currently Implemented Method: Do Prospective Improvements Outweigh Constraints and Costs?

In a few cases (e.g., entirely new surveys), one may have open a wide range of options for development and implementation of outlier methods. In most cases that involve ongoing surveys, however, any proposal to change a currently implemented outlier procedure would require consideration of three questions.

- What is the performance of the proposed new procedure, relative to the old procedure, as measured by mean squared error and other measures of risk considered in Section 3.2?

- What operational constraints, if any, would need to be changed to allow implementation of the proposed new procedure?

- Do the anticipated improvement in performance cross a sufficiently high threshold to justify the costs and modification of operational constraints required to implement the proposed new procedure?

If would be of interest to expand the current mathematical statistics literature to account for these questions in a systematic form.

4.6. Evaluation of Performance in the Presence of Gross Measurement Error

As noted in Section 1.1, the formal literature on outliers and influential observations tends to focus primarily on observations that arise from "true values" but that are influential due to a large value or a large weight, or both. In addition, agency documents sometimes state the assumption that a previous edit step has identified and corrected any gross errors that may have been in the data.

However, some users of outlier-detection tools indicate that applications of these tools do sometimes lead to identification of gross measurement errors. Thus, as a supplement to traditional evaluations of outlier-detection procedures, it would be worthwhile to develop evaluations of bias, variance and other properties with respect to two dimensions of random variability:

- the dimension of variability generally considered in the outlier literature, e.g., induced by a sample-selection mechanism or by an underlying superpopulation model for the "true values" ; and

- an additional dimension of variability induced by a (random) measurement-error process. This second component might be captured by a two-component mixture model in which the second component equals zero with probability $(1-\varepsilon)$, and follows a severely skewed distribution with probability ε, where ε is a very small number.

Under the abovementioned two dimensions of variability, an important question is whether certain procedures are not optimal under the "pure" design dimension, but are more robust against the gross-error contamination under the second error-model dimension.

4.7. Robust Modeling for Genralized Variance Functions and Components of Design Effects; and for Small Domain Estimation

This paper focuses on outlier issues in estimation and inference for descriptive estimands of large population aggregates. There is also a large body of research on outliers and influential observations in the literature on statistical modeling. We will not attempt to cover this material in depth, but we note two important sample survey areas to which the modeling work might be extended.

First, statistical agencies often develop models for generalized variance functions and for the components of design effects. See, e.g., Valliant (1987), Johnson and King (1987), United Nations Statistical Division (2005) and references cited therein. Much of this work involves the use of regression models and other generalized linear models to describe the relationship between sample variances (or transformations thereof) and a prespecified set of preditor variables. For cases in which some units are very large, the associated sample variances can be unstable, which can in turn lead to problems with the performance of standard estimation methods for generalized variance function models. Similar issues have been noted in a partly related literature on variance-function modeling in biometrics and engineering applications, e.g., Giltinian et al. (1986), Davidian and Carroll (1987, 1988) and Carroll and Welsh (1988). It would be of interest to study extensions of the robust variance-function modeling literature to work with generalized variance functions and design effects.

Second, there has been increasing interest in the use of models to produce estimators for small domains, i.e., subpopulations from which a given survey has collected a relatively small number of observations. For some general background on small area estimation, see, e.g., Rao (2003), Nandram and Sedransk (2002), Malec et al. (1997), Wolter and Causey (1991), Battese et al. (1988), Platek et al. (1987) and references cited therein. In small area work, it is often noted that some observations can be influential for estimation of a mean or total in a small area or small domain, even if the same observation is not influential for estimation of quantities at the national or large-subpopulation level. This illustrates the idea that the status of an observation as influential always depends on the specific estimand and point estimator being considered. In addition, small domain estimation methods generally involve the use of linear or hierarchical models for the relationship between unit or area-level outcome variables and a set of available auxiliary variables (obtained, e.g., through administrative records). Consequently, many of the issues, diagnostics and estimation techniques that arise with influential observations in linear and hierarchical models carry over to small domain estimation settings.

5. Questions for FESAC on "Outliers and Influential Observations in Establishment Surveys"

For each of the topics listed below, we are seeking input from FESAC members on the following questions.

1. Beyond the references cited in this paper, is there specific literature that we should consider in developing and evaluating appropriate methodology to address these issues?

2. Are there specific empirical cases that you have encountered that would shed light on these issues?

3. In light of the literature and your previous experience in this area, what are the predominant factors that we should consider in developing and implementing additional methodology to address these issues?

5.1 Mathematical Statistics

5.1.1. What do you consider some methods that strike an appropriate balance between addressing the effects of outliers, without masking important information that they may convey?

5.1.1. What do you consider appropriate methods for identification and treatment of outliers when agencies prefer to use a single unified (edited) set of data and weights to produce estimates at national, state and MSA levels (and as elementary estimators for small domain estimation)?

5.1.2. What do you consider appropriate methods for multivariate outlier identification and treatment, especially in light of high-dimensional utility functions of multiple stakeholders?

5.2. Behavioral Sciences and Computer Sciences

5.2.1. In visual displays intended for outlier detection, what do people tend to see or miss? Also, are there specific methods we should use to reduce the risk of visual reviews missing or misinterpreting outliers?

5.2.2. Implementation or enhancement of some outlier-detection ideas may potentially involve concepts or tools from the literature on artificial intelligence (defined broadly to include machine learning and adaptive models). Please comment on:

- Previous artificial-intelligence-type methods used in (or considered for) outlier detection, and any special issues encountered in the development or implementation of such methods. In general, we would seek tools that would, within specified areas, help us capture analysts' insights;

provide tools that help analysts use their insights more efficiently; and capture additional information that analysts currently may not be using.

- Specific areas of outlier-detection work in which you think artificial intelligence methods could potentially produce substantial improvements, relative to currently available tools

5.3. Economics

5.3.1. For purposes of outlier detection and treatment, agencies are sometimes told that "local economic knowledge" can provide substantial value added, relative to the information available to national-level analysts

- Are there specific tools that should be used to evaluate the incremental value added from "local economic knowledge"?

- Have there been specific empirical studies of this incremental value added? If so, what were the results?

5.3.2. Please comment on any experiences you have had with systematic gross errors in administrative record systems attributable to definitional differences, inconsistent aggregation methods, or other sources.

5.3.3. Please comment on specific economic models (especially for establishment-level pheomena) that identify predictors of extreme observations (relative or absolute)

5.3.4. Please comment on specific extreme-value or mixture-distribution models that have been successfully applied to establishment-level data

5.4. Systems Development and Integration

Please comment on any generalized data-processing system modules for outlier treatment and detection that may have been developed at your institution or other institutions. What are the primary efficiency gains or losses attributable to use of generalized modules, relative to case-specific methods, possibly including:

- Reduction of personnel costs through generalized module approaches

- Other efficiencies of scale

- Losses attributable to a general module not adequately meeting the perceived needs of individual programs (or not meeting these needs without extensive tailoring that cancels the efficiencies of scale anticipated above).

Acknowledgements

The authors thank Terry Burdette, Larry Ernst, Julie Gershunskaya, Howard Hogan, Erin Huband, Larry Huff and Mary Mulry for helpful discussions. The second half of Section 3.5 was written by Mary Mulry.

This paper has been prepared for presentation to the Federal Economic Statistics Advisory Committee (FESAC) on June 9, 2006. It represents work in progress and does not represent any agency's final positions on issues addressed. The FESAC is a Federal Advisory Committee sponsored jointly by the Bureau of Labor Statistics of the U.S. Department of Labor and by the Bureau of Economic Analysis and the Bureau of the Census of the U.S. Department of Commerce.

References

Anderson, A.E., S.H. Cohen, E. Murphy, E. Nichols, R. Sigman and D.K. Willimack (2003). Changes to Editing Strategies When Establishment Survey Data Collection Moves to the Web. Paper presented to the Federal Economic Statistics Advisory Committee, March 21, 2003. http://www.bls.gov/bls/fesacp2032103.pdf

Barkume, A.J. and T.G. Moehrle (2004). Development of an ECI Excluding Workers Earning Incentive Pay. Paper presented to the December 14, 2004 meeting of the Federal Economic Statistics Advisory Committee.

Barsky, C., A. Dorfman, K. Dwork-Fisher and L. Ernst (2003). Report of the OCLT Data Outlier Team, Phase 1. Internal memorandum, Bureau of Labor Statistics.

Barsky, C., A. Dorfman, K. Dworak-Fisher, L. Ernst and C. Guciardo (2003). Report of the OCLT Data Outlier Team, Phase 2. Internal memorandum, Bureau of Labor Statistics.

Battese, G.E., R.M. Harter and W.A. Fuller (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data' *Journal of the American Statistical Association* **83**, 28-36

Beaumont, J.-F. (2004). Robust Estimation of a Finite Population Total in the Presence of Influential Units. Report for the U.K. Office of National Statistics, Newport, Wales.

Beaumont, J.-F. and A. Alavi (2004). Robust Generalized Regression Estimation. *Survey Methodology* **30**, 195-208.

Beguin, C. and B. Hulliger (2004). Multivariate Outlier Detection in Incomplete Survey Data: The Epidemic Algorithm and Transformed Rank Correlations. *Journal of the Royal Statistical Society, Series A* **167**, 275-294.

Bibby, J. and H. Toutenberg (1977). *Prediction and Improved Estimation in Linear Models*. New York: Wiley.

Bienias, J.L. (1995). Methods for Outlier Detection for the Quarterly Financial Report. Economic Statistical Methods Report Series ESMD-9407, Bureau of the Census.

Bienias, J.L., T. Bell, W. Caldwell, V. Garrett, I. Hall, H. Hogan, D. Hundertmark, J. Juzwiak, W. Knowlton, D. Lassman, S. Scheleur, D. Stachurski and G. Wright (1994). Graphical Review of Economic Data. Economic Statistical Methods Division Report Series ESMD-9406, Bureau of the Census.

Carroll, R.J. and A.H. Welsh (1988). A Note on Asymmetry and Robustness in Linear Regression. *The American Statistician* **42**, 285-287.

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association* **81**, 1063-1069.

Chambers, R.L. (1996). Robust Case-Weighting for Multipurpose Establishment Surveys. *Journal of Official Statistics* **12**, 3-32.

Chambers, R.L., A.H. Dorfman and T.E. Wehrly (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. *Journal of the American Statistical Association* **88**, 268-277.

Chambers, R., A. Hentges and X. Zhao (2004). Robust Automatic Methods for Outlier and Error Detection. *Journal of the Royal Statistical Society, Series A* **167**, 323-339.

Chambers, R., P. Kokic, P. Smigh and M. Cruddas (2000). Winsorization for Identifying and Treating Outliers in Business Surveys. *Proceedings of the Second International Conference on Establishment Surveys*, 717-726. Alexandria, Virginia: American Statistical Association.

Chambers, R. L. and Ren, R. (2004). "Outlier Robust Imputation of Survey Data." *2004 Proceedings of the American Statistical Association.*

Clarke, M. (1995). "Winsorization Methods in Sample Surveys." Masters Thesis. Department of Statistics. Australia National University.

Davidian, M. and R.J. Carroll (1987). Variance Function Estimation. *Journal of the American Statsitical Association* **82**, 1079-1091.

Davidian, M., R.J. Carroll and W. Smith (1988). Variance Functions and the Minimum Detectable Concentration in Assays. *Biometrika* **75**, 549-556.

Detlefsen, R. (1992). A Weight Adjustment Technique. Internal Memorandum, Bureau of the Census.

DiZio, M., O. Luzi and A. Manzari (2005). Evaluating Editing and Imputation Processes: The Italian Experience. *Proceedings of the Work Session on Statistical Data Editing, United Statistical Commission and Economic Commission for Europe.*

Elliott, M.R. and R.J.A. Little (2000). Model-Based Alternatives to Trimming Survey Weights. *Journal of Official Statistics* **16**, 191-209.

Ernst, L. (1980). Comparison of Estimators of the Mean which Adjust for Large Observations. *Sankhya, Series C* **42**, 1-16.

Esposito, R., J.K. Fox, D. Lin and K. Tidemann (1994). ARIES: A Visual Path in the Investigation of Statistical Data. *Journal of Computational and Graphical Statistics* **3**, 113-125.

Francisco, C.A. and W.A. Fuller (1991). Quantile Estimation with a Complex Survey Design, *The Annals of Statistics* **19**, 454-469

Franklin, S., S. Thomas and M. Brodeur (2000). Robust Multivariate Outlier Detection Using Mahalanobis' Distance and Modified Stahel-Donoho Estimators. *Proceedings of the Second International Conference on Establishment Surveys*, 697-706. Alexandria, Virginia: American Statistical Association.

Fuller, W.A. Simple Estimators for the Mean of Skewed Populations. *Statistica Sinica* **1**, 137-158.

Gershunskaya, J. and L. Huff (2004). Outlier Detection and Treatment in the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Giltinan, D.M., R.J. Carroll and D. Ruppert (1986). Some New Estimation Methods for Weighted Regression When There Are Possible Outliers. *Technometrics* **28**, 219-230.

Gomish, J.M. and R.S. Sigman (2003). Improving the Detection and Treatment of Outliers in the Consumer Expenditure (CE) Survey and the Survey of Residential Alterations and Repairs (SORAR). Economic Statistical Methods Report Series ESM-0301, Bureau of the Census.

Gwet, J.-P. and H. Lee (2000). An Evaluation of Outlier-Resistant Procedures in Establishment Surveys. *Proceedings of the Second International Conference on Establishment Surveys*, 707-716. Alexandria, Virginia: American Statistical Association.

Gwet, J.-P., and L.-P. Rivest (1992). Outlier Resistant Alternatives to the Ratio Estimator. *Journal of the American Statistical Association* **87**, 1174-1182.

Hartley, H.O. and J.N.K Rao (1968). A New Estimation Theory for Sample Surveys. *Biometrika* **55** 547-557.

Hedlin, D., H. Falvey, R. Chambers and P. Kokic (2001). Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics* **17**, 527-544.

Hidiroglou, M.A. and J.-M. Berthelot (1986). Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology* **12**, 73-83

Hidiroglou, M.A. and K.P. Srinath (1981). Some Estimators of a Population Total from Simple Random Samples Containing Large Units. *Journal of the American Statistical Association* **76**, 690-695.

Huber, P. J. (1964) "Robust Estimation of a location parameter". *Annals of Mathematical Statistics*. Institute of Mathematical Statistics. 35. 73-101.

Hulliger, B. (1995). Outlier Robust Horvitz-Thompson Estimators. *Survey Methodology* **21**, 79-87.

Hulliger, B. (1999). Simple and Robust Estimators for Sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 54-63.

Hunt, J.W., J.S. Johnson and C.S. King (1999). Detecting Outliers in the Monthly Retail Trade Survey Using the Hidiroglou-Berthelot Method. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 539-547.

Johnson, E.G. and B.F. King (1987). Generalized Variance Functions for Complex Sample Survey. *Journal of Official Statistics* **3**, 235-250.

Kokic, P.N. and P.A. Bell (1994). Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. *Journal of Official Statistics* **10**, 419-435.

Latouche, M. and J.-M. Berthelot (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics* **8**, 389-400.

Lawrence, D. and C. McDavitt (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics* **10**, 437-447.

Lee, H. (1991). Model-Based Estimators That Are Robust to Outliers. *Proceedings of the 1991 Annual Research Conference, Bureau of the Census* 178-202. Washington, DC: U.S. Department of Commerce.

Lee, H. (1995). Outliers in Business Surveys. Chapter 26 in *Business Survey Methods* (B. Cox et al., eds.) New York: Wiley.

Little, R.J.A. and P.J. Smith (1987). Editing and Imputation for Quantitative Survey Data. *Journal of the American Statistical Association* **82**, 58-68.

Lutz, S.M., V.A. McCracken and R.C. Mittlehammer (1991). Preparing Large Survey Samples for Empirical Analysis: Outlier Detection Using Robust Versus Non-Robust Estimators. *Proceedings of the 1991 Annual Research Conference, Bureau of the Census* 157-177. Washington, DC: U.S. Department of Commerce.

Luzi and A. Pallara (1999). Combining Macroediting and Selective Editing to Detect Influential Observations in Cross-Sectional Survey Data. *Proceedings of the Work Session on Statistical Data Editing, United Statistical Commission and Economic Commission for Europe.*

Malec, D., J. Sedransk, C.L. Moriarity and F.B. LeClere (1997). Small Area Inference for Binary Variables in the National Health Interview Survey, *Journal of the American Statistical Association* **92**, 815-826

McConnell, S. and W. Goodman (1994). Recognition of More than One Possible Trend in Time Series: Redesigned Screening of Microdata in the Current Employment Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Nandram, B. and J. Sedransk (2002). Bayesian Predictive Inference for the Proportion of Eroded Land in a Small Area in Iowa, *Environmental and Ecological Statistics*, **9**, 221-236.

Platek, R., J.N.K. Rao, C.-E. Sarndal and M.P. Singh (eds.) (1987). *Small Area Statistics.* New York: Wiley.

Potter, F.J. (1988). Survey of Procedures to Control Extreme Sampling Weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 453-458.

Potter, F.J. (1993). The Effect of Weight Trimming on Nonlinear Survey Estimates. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 758-763.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.

Rao, J.N.K., J.G. Kovar and H.J. Mantel (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information, *Biometrika* **77**, 365-375

Rivest, L.-P. and D. Hurtubise (1995). On Searls' Winsorized Mean for Skewed Populations. *Survey Methodology* **21**, 107-116.

Rubin-Bleuer, S. and the Workplace and Employee Survey Team (2003). Outlier Detection in Workplace and Employee Survey (WES). Internal memorandum, Statistics Canada.

Saidi, A. and S. Rubin-Bleuer (2005). Detection of Outliers in the Canadian Consumer Price Index. *Proceedings of the Work Session on Statistical Data Editing, United Statistical Commission and Economic Commission for Europe.*

Scott, S., A. Clinton, H. Katkki, J. Buszuwski, C. Ponikowski, T. Erickson and D. Swanson (1999). A Study of the Handling of Outliers: Preliminary Report of the BLS Outlier Team. Internal Report, Bureau of Labor Statistics.

Searls, D.T. (1966). An Estimator for a Population Mean Which Reduces the Effect of Large True Observations. *Journal of the American Statistical Association* **61** 1200-1204.

Smith, T.M.F. (1987). Influential Observations in Survey Sampling. *Journal of Applied Statistics* **14**, 143-152.

Smith, T.M.F., and E. Njenga (1992). Robust Model-Based Methods for Analytic Surveys. *Survey Methodology* **18**, 187-208.

Theberge, A. (2000). Calibration and Restricted Weights. *Survey Methodology* **26** 99-107.

Tikku. M.L. (1983). Exact Efficiencies of Some Robust Estimators in Sample Survey. *Communications in Statistics, Theory and Methods* **12**, 2043-2051.

United Nations Statistical Division (2005). *Household Sample Surveys in Developing and Transition Countries.* New York: United Nations. Available at: http://unstats.un.org/unsd/hhsurveys/

U.S. Department of Labor (2003). *Estimation System Procedures Manual, Division of Safety and Health Systems, Bureau of Labor Statistics.*

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association* **82**, 499-508.

Welsh, A.H. and E. Ronchetti (1998). Bias-Calibrated Estimation from Sample Surveys Containing Outliers. *Journal of the Royal Statistical Society, Series B* **60**, 413-428.

Wolter, K.M. (1995). *Introduction to Variance Estimation*. New York: Springer Verlag.

Wolter, K.M. (1996a). Specifications for Handling Large Observations. Memorandum to the Bureau of Labor Statistics, April 19, 1996.

Wolter, K.M. (1996b). Testing the Outlier Procedure Using Illinois Data on Service Industries. Memorandum to the Bureau of Labor Statistics, June 6, 1996.

Wolter, K.M. and B.D. Causey (1991). Evaluation of Procedures for Improving Population Estimates for Small Areas, *Journal of the American Statistical Association* **86**, 278-284.

Zaslavsky, A.M., N. Schenker and T.R. Belin (2001). Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey.